This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.

第 28 回日本バーチャルリアリティ学会大会論文集（2023 年 9 月）

Neural Radnaice Fields による人や車を除いた自由視点映像合成

Free-viewpoint video synthesis removing people and cars with Neural Radnaice Fields

大隣嵩 [1)]，池畑諭 [2),3),1)]，相澤清晴 [1)]
Takashi OTONARI, Satoshi IKEHATA, and Kiyoharu AIZAWA

1) 東京大学 （〒 113-0033 東京都文京区本郷 7-3-1）
2) 国立情報学研究所（〒 101-8430 東京都千代田区一ツ橋 2-1-2)
3) 東京工業大学（〒 152-8550 東京都目黒区大岡山 2-12-1)

概要: 本論文では，人や車などの動く物体を含む動画から，動く物体を除いた自由視点映像を作成する課題に取り組む．この課題を達成するために，Neural Radnaice Fields を用い，動く物体をマスクしたセグメンテーションマスクを組み合わせた．結果として，頭より上に構えた 360 度カメラを使って撮影した 360 度映像から，カメラの高さを変えた猫から見た視点や人間の視点などの映像を作成することができた．

キーワード： 自由視点映像, 静動分離, セマンティックセグメンテーション

1. はじめに

Google Maps の immersive view [1] や，スポーツ中継における自由視点映像 [2] のように，複数視点の画像から，シーンの任意視点からの画像を合成する任意視点画像合成（Novel View Synthesis: 以下 NVS）が実用化されつつある．この分野における最近のブレークスルーである Neural Radiance Fields（以下 NeRF） [3] では，カメラパラメータを求めた多視点画像群から静的シーンの NVS を行うことができる．しかし，実世界シーンの撮影では，動く物体が含まれることが多く，シーンから一時的に映った物体の除去や，静的シーンを背景としたビジュアルエフェクト（Visual Effects: VFX）などのコンテンツ作成において，動く物体を除去しながら静的シーンのみの NeRF 表現を学習することが望まれている．オリジナルの NeRF で動く物体を含む動画を学習に用いると，動く物体が浮遊物（floaters）として表現され，静的シーンの学習に失敗してしまう [4].

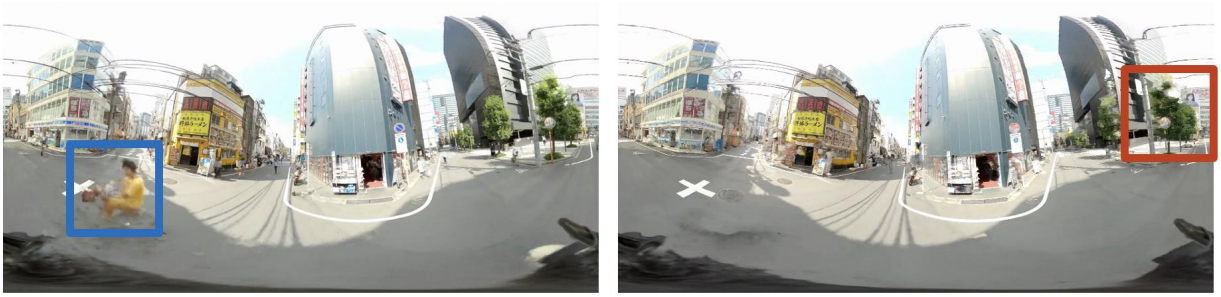
RobustNeRF [4] では，動く物体の残差は常に大きくなり，画像の大きくつながった領域を占めるという帰納バイアスを利用して動く物体となるアウトライアを判別・除去する方法を提案している. この手法は静動分離において高い性能を示したが，人や車などの運動物体が画像内で占める領域が小さい実世界のシーンではうまく機能しない．これは，アウトライアが占める領域が小さい場合，RobustNeRF が提案したアウトライアに関する帰納バイアスが弱くなるためである．その結果，アウトライアのサイズを決めるパッチサイズをデフォルトの値に設定するとアウトライアを除去することができず，パッチサイズをアウトライアのサイズに合わせて小さく設定し，アウトライアの帰納バイアスを弱くするとインライアを誤って除去することが多くなる（図 1）.

本論文では，既存の研究では取り組まれなかった，動く小さな物体が多く存在する実世界のシーンで静動分離を行う課題に取り組む．そのために，動く物体が画像内の小さな領域を占める場合にアウトライアに関する帰納バイアスが弱くなり，インライアと区別することが難しくなるという RobustNeRF が抱える問題に対するシンプルな解決策としてアウトライアとなる動く物体のラベルを事前知識として利用したセグメンテーションマスクを導入する．さらに，2 次元画像のセグメンテーションモデルによるエラーを軽減するために，Test Time Augmentation と学習時におけるマスク画像の膨張処理を用いた，本タスクに適したセグメンテーションマスクの改良方法を提案した．実世界のシーンを用いた実験により，RobustNeRF 単体では，分離が失敗するシーンにおいて，追加で与えられるセグメンテーションマスクにより効果的に動く物体を除去できることを視覚的に示した．また，学習した NeRF は視点を自由に移動することが可能で，カメラの高さを変えた猫から見た視点や人間の視点などの映像が作成できることを示した.

2. RobustNeRF

RobustNeRF [4] は，動く物体を含む画像群から，静的な背景シーンのみを取り出すために，動く物体をアウトライアとして取り除く robust loss を導入した．光線 $\mathbf{r}$ に対する robust loss $\mathcal{L}^{\mathbf{r}}_{\text{robust}}$ は，式 (1) のように定義される.

© 2023 日本バーチャルリアリティ学会



デフォルトのパッチサイズ　　動く物体を除去するようにパッチサイズを調整

図 1: RobustNeRF ではパッチサイズより小さなサイズのアウトライアを除去することができずデフォルトのパラメータ設定では左図の青枠内の動く物体を除去できない．一方で，動く物体を漏れなく除去するようにパラメータを調整するとシーンの詳細が学習できないとともに，右図の赤枠内のように背景として学習する必要のある木が適切に学習されない．左図は式 (2) において 90 パーセンタイル以下の残差をインライアとし式 (3) のパッチサイズを 16×16 とした結果．右図は式 (2) において 0.8 パーセンタイル以下の残差をインライアとし式 (3) のパッチサイズを 4×4 とした結果．

$$\mathcal{L}^{\mathbf{r}}_{\text{robust}} = \omega(\mathcal{R}_8(\mathbf{r}))\|C_{\text{pred}}(\mathbf{r}) - C_{\text{gt}}(\mathbf{r})\|_2^2 \quad (1)$$

合成した画像の色 C_{pred} と正解の色 C_{gt} から計算されるオリジナルの NeRF で使う再構成誤差に，光線 $\mathbf{r}$ の周囲 8×8 のパッチの光線 $\mathcal{R}_8(\mathbf{r})$ を入力とする重み関数 $\omega(\mathcal{R}_8(\mathbf{r}))$ を追加し，アウトライアの重みを 0 にすることで学習から除外する．アウトライアとなる動く物体は学習が難しいため残差が常に大きくなり，画像の大きくつながった領域を占めると考えられる．そのため，残差の大きなピクセルの空間的な滑らかさを捉えることでアウトライアを判別する．また，インライアの残差の大きなピクセルはスパースに存在すると仮定し，残差の大きなピクセルが大きくつながった領域を占めるアウトライアと区別している．以上の帰納バイアスを基に，robust loss では，パッチで与えられる周囲の光線（ピクセル）の情報から，アウトライアを判定し，学習から除外する．

robust loss の重み関数 ω の計算では，まずイテレーションごとの残差の相対的な大きさを定めるために，光線 $\mathbf{r}$ ごとに求めた残差 $\epsilon(\mathbf{r})$ のソートを行い，あるパーセンタイル以下の残差はインライアであると仮定する．式 (2) では中央値以下の残差を持つピクセル（光線）をインライアと仮定している．

$$\tilde{\omega}(\mathbf{r}) = \epsilon(\mathbf{r}) \leq \mathcal{T}_\epsilon, \mathcal{T}_\epsilon = \text{Median}_{\mathbf{r}}\{\epsilon(\mathbf{r})\} \quad (2)$$

次に，アウトライアの空間的な滑らかさを捉えるために，式 (3) のように各値が $\frac{1}{9}$ で 3×3 のカーネル $\mathcal{B}_{3\times3}$ により畳み込み演算を行い，その値が 0.5 以上の場合，インライアとする．つまり，式 (2) でインライアと仮定したピクセルが周囲に半分以上の割合で存在する場合，その点をインライアとする．これにより，空間的にスパースに存在する残差の大きなインライアがアウトライアとして分類されることがなくなる．

$$\mathcal{W}(\mathbf{r}) = (\tilde{\omega}(\mathbf{r}) \circledast \mathcal{B}_{3\times3}) \geq \mathcal{T}_\circledast, \mathcal{T}_\circledast = 0.5 \quad (3)$$

式 (3) のみでは，細かいテクスチャを持つインライアをアウトライアと誤分類することが多くなるため，さらに，アウトライアの空間的な滑らかさへの帰納バイアスを強くする．式 (4) では，式 (3) で求めたインライアの割合が周囲 16×16 のパッチ内で 0.6 以上の時，インライアとする．

$$\omega(\mathcal{R}_8(\mathbf{r})) = \mathbb{E}_{\mathbf{s}\sim\mathcal{R}_{16}(\mathbf{r})}[\mathcal{W}(\mathbf{s})] \geq \mathcal{T}_\mathcal{R}, \mathcal{T}_\mathcal{R} = 0.6 \quad (4)$$

以上の式 (2)，式 (3)，式 (4) により，アウトライアとなる動く物体が残差が大きく，画像の大きくつながった領域を占めるという帰納バイアスを捉え，インライアとアウトライアの判別を行うことが可能となる．

3. 静動分離に適したセグメンテーションマスクの作成

RobustNeRF では，動く物体の種類に関する事前知識を前提とせずに，残差の空間的な分布からアウトライアを判別・除去する手法を提案した．しかし，図 1 に示したように，実世界のシーンでしばしば起こる，動く物体が画像内の小さな領域を占める場合にアウトライアに関する帰納バイアスが弱くなり，インライアと区別することが難しくなる．この問題に対するシンプルな解決策として，本論文では，アウトライアとなる動く物体のラベルを事前知識として利用したセグメンテーションマスクを導入する．セグメンテーションマスクを追加することにより，事前に定義した動く物体が静止している場合も NeRF の学習から除外することになる．本論文では，より良いセグメンテーションマスクを得るための方法として (1) Test Time Augmentation による方法，(2) セグメンテーションマスクを膨張させる方法の 2 つのセグメンテーションマスクの改善手法を導入する．

Test Time Augmentation を用いたセグメンテーションマスクの改良: まず，画像認識やセグメンテーションで推論時の精度を上げるために使われる Test Time Augmentation [5, 6] をタスクに合った形で拡張した．図 2 のように，5 種類のオーグメンテーション（contrast, horizontal flip, median blur, gamma, sharpen）を加えた画像と，オリジナルの画像を入力してセグメンテーションマスクを取得し，各ピクセルについて 1 枚以上の画像で動的成分と判定すれ

© 2023 日本バーチャルリアリティ学会

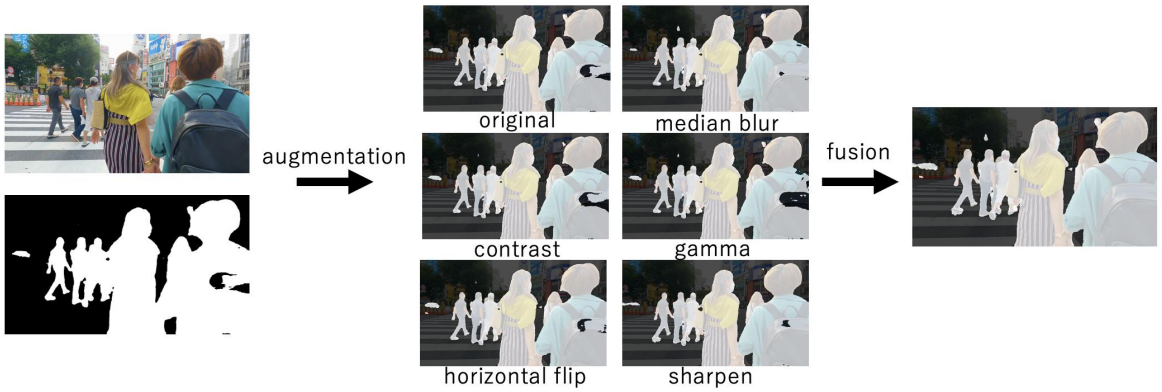


図 2: Test Time Augmentation によるセグメンテーションマスクの改良．5 種類のオーグメンテーションを行った画像とオリジナルの画像からセグメンテーションマスクを取得し，各ピクセルについて 1 枚以上の画像で動的成分となった場合は動的成分とするように，セグメンテーションマスクを融合する．

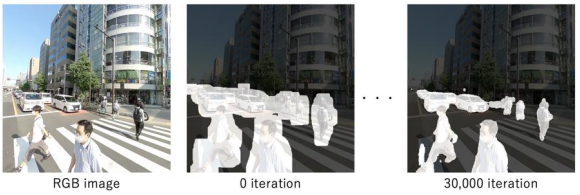


図 3: 学習が不安定になる初期段階では，動く物体をマスクしないようにマスクの膨張処理を行う．学習の進行とともに RobustNeRF により動く物体の判定が正確になることが期待されるため，最終段階のマスクでは膨張処理を行わない．

ば，動的成分とする方法でセグメンテーションマスクを融合する．以上により，動的成分に関するセグメンテーション漏れが軽減される．

セグメンテーションマスクの膨張: 次にセグメンテーションマスクの膨張処理により，学習の安定化を行う．学習が不安定になる学習初期にセグメンテーションマスクのエラーが混在すると，動く物体を誤って静的な背景と判定する可能性が高くなる．そのため，最初は膨張処理を行ったセグメンテーション漏れの少ないマスクを入力し，徐々にオリジナルのマスクになるよう収縮させることで，物体境界付近の静的な背景シーンを学習に含めつつ，RobustNeRF における学習初期の動く物体の判定エラーを軽減させる．動的成分を 1，静的成分を 0 としたセグメンテーションマスクの膨張は，値が 1 でサイズ 3×3 のカーネルで畳み込みを行い，値が 1 以上の場合に膨張後のセグメンテーションマスクが 1 となるように膨張させる．実際に，セグメンテーションマスクを膨張させると図 3 のように学習初期にはセグメンテーションマスクの動的成分が広くなる．

4. 実験

本論文で導入したセグメンテーションマスクを RobustNeRF [4] に追加した効果を検証するために実験を行った．ベースラインとして nerfstudio [7] が提供する nerfacto を使用し，nerfacto 上で実装した robust loss を元に，セグメンテーションマスクや robust loss の有無による視覚的な比較を行った．robust loss を用いない学習では 1 バッチあたり 16,896 本の光線を使用し，robust loss を用いる場合は，サイズが 16×16 の 66 個のパッチを 1 バッチの学習に用いた．学習は，NVIDIA A100 GPU で 30,000 イテレーション（約 30 分）を要した．セグメンテーションマスクは，ADE20K [8] で学習した SegFormer [9] により取得し，person，car，bus，truck，van，minibike，animal，bicycle を動く物体のラベルと定義した．セグメンテーションマスクの膨張処理では，2,000 イテレーションごとにイテレーション数を i とし，オリジナルのマスクに対して $14 - \frac{i}{2000}$ 回の膨張処理を行った．また，全ての実験を通して式 (2) の残差のソートでは，0.9 パーセンタイル以下の残差をインライアと仮定した．

4.1 実世界シーンを用いた定性的な比較

秋葉原で撮影した 3 つの 360 度動画を用いて RobustNeRF と提案手法の定性的な比較を行った結果を図 4 に示す．この実験では，360 度動画の各フレームから視野角 90 度，画像サイズ 512×512 で 14 枚ずつ透視投影画像を切り出し，COLMAP によりカメラパラメータを求めた．図 4 のように，RobustNeRF では分離が失敗するケースでも，セグメンテーションマスクを追加することにより，インライアとアウトライアの分離ができるようになっている．この結果は，動く物体を含む画像群からの静的シーンの復元において，RobustNeRF にセグメンテーションマスクを組み合わせることが有効であることを示している．

次に，撮影時から視点を変更し，人の視点，猫の視点から画像合成した結果を図 5 に示す. 図 5 より，人の視点，猫の視点から違和感の少ない画像を合成しつつ，学習時に存在していた人や車などの動く物体を消去することができた．撮影時の人の頭より高いカメラ位置から人の視点まで下げることでコンテンツへの没入感をより高めることが期待でき，また，猫の視点では普段体験できないより低い視点を与える映像を作成することができる．

© 2023 日本バーチャルリアリティ学会

Reference image　nerfacto + segmentation mask　nerfacto (robust loss)　nerfacto (robust loss) + segmentation mask

図 4: 秋葉原で撮影したシーンによる定性的な比較．nerfacto の学習では，カメラパラメータの最適化も行っているため，学習画像群のうちレンダリングした画像の視点から近い画像を Reference image として選択した．

Reference image　人の視点　猫の視点

図 5: 視点の高さを変更し，透視投影画像を合成した結果．

5. まとめ

本研究では Test Time Augmentation により取得したセグメンテーションマスクを追加することで，既存手法では適用が難しかった動く小さな物体が多く映る実世界のシーンに対して，効果的に動く物体が取り除かれることを実験を通して示した．秋葉原で撮影した 360 度動画を用いた実験により，学習時の視点とは異なる，カメラの高さを変えた猫の視点や人間の視点などの映像が違和感なく作成できることを実証した．

謝辞

本研究の一部は、JST-Mirai Program JPMJMI21H1、JSPS KAKENHI 21H03460 の支援を受けた

参考文献

[1] Google Maps. https://www.google.com/maps.

[2] Canon Volumetric Video. https://global.canon/en/vvs/works/works-live-baseball.html.

[3] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[4] Daniel Duckworth Ivan Krasin David J. Fleet Sara Sabour, Suhani Vora and Andrea Tagliasacchi. Robustnerf: Ignoring distractors with robust losses. *arXiv:2302.00833*, 2023.

[5] Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *MIDL*, 2018.

[6] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *CVPR*, 2022.

[7] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. *arXiv:2302.04264*, 2023.

[8] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.

[9] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021.

© 2023 日本バーチャルリアリティ学会